ML Beyond Curve Fitting: An Intro to Causal Inference and do-Calculus

Reference: https://www.inference.vc/untitled/

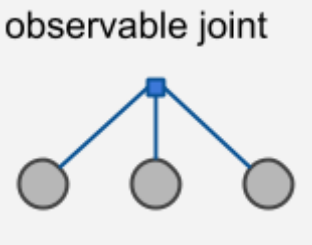
- Correlational models and statistical learning methods essentially focuses of finding the conditional probability $P(Y|X)$ where Y is the outcome variable and X is the predictor. Both X and Y are RVs.
- $P(Y|X) = \frac{P(X,Y)}{P(X)}$. Note that I can compute the marginal probability $P(X)$ from the joint probability $P(X,Y)$. $P(X,Y)$ and $P(X)$ can be estimated from the train data (recall Naïve Bayes).
- In general, our train dataset is sampled from a random process $P(X,Y,\dots)$. The goal of statistical learning is to estimate $P(X,Y\dots.)$, from which $P(Y|X)$ can be obtained.

Distinction between $p(y|X)$ and $p(y|doX)$

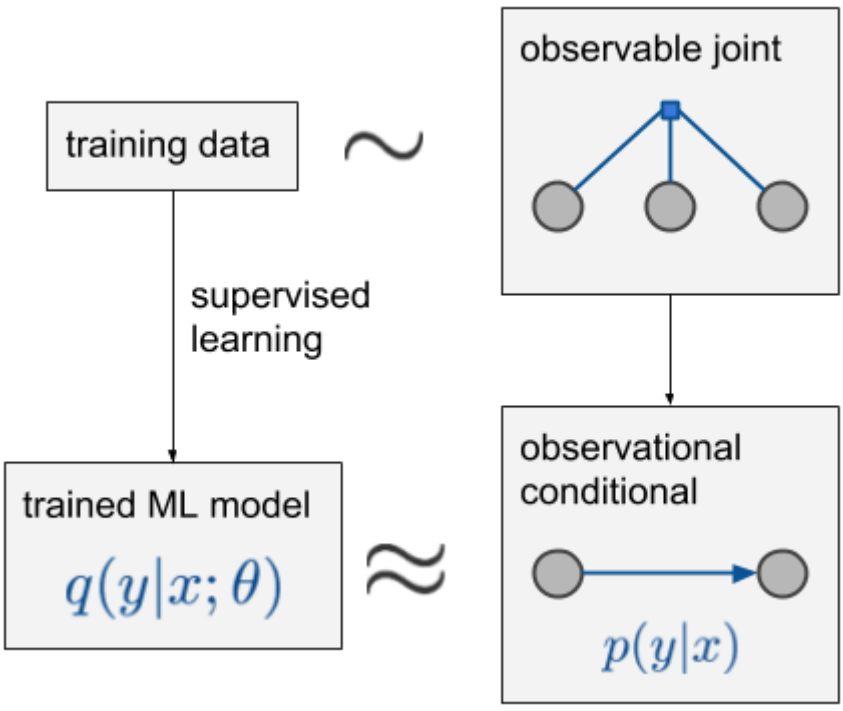
- Interventional vs observational conditionals.
- $p(Y|X=x)$: distribution Y given that I observe X=x
- p(Y|do(X=x)): distribution of Y if I were to set value of X to x.
- In general $p(Y|X=x) \neq p(Y|do(X=x))$.
    - The distinction can be illustrated by the expresso machine pressure example. If I artificially intervene and forces the barometer reading to x (do x), then this won't affect the distribution of Y.
    - In contrast, knowing "observational" information of X (barometer readings), it tells me something about the the probability of Y:
        - $P(Y|X) \neq P(Y)$ since X provides "new information" about the actual pressure of the expresso machine. X and Y are dependent. do(X) and Y are NOT dependent.
- p(y|X) is estimated for applications that involve forecasts, p(y|do(X)) is estimated for causal inference (ex: impact evaluation of X).
    - Look at this another way, the estimation of $p(y|do(X))$, will "inform" how I choose X=x to achieve some desirable outcome $Y=y$.
- I can obtain p(Y|do(x=1)) from $P_{do(X=x)}(x,z,y)$.
- $P_{do(X=x)}(x,z,y)$ is the joint probability of my outcome and predictors that we would observed had we carry out the intervention (force X=x). It is different from $P(x,z,y)$.
- Using quasi-experimental methods or Bayesian networks, we are actually trying to estimate $P(y|do(X))$ or $P_{do(X=x)}(x,z,y)$. If we can perform RCT, then we directly measure $P(y|do(X)) = ATET$.

Causal Models

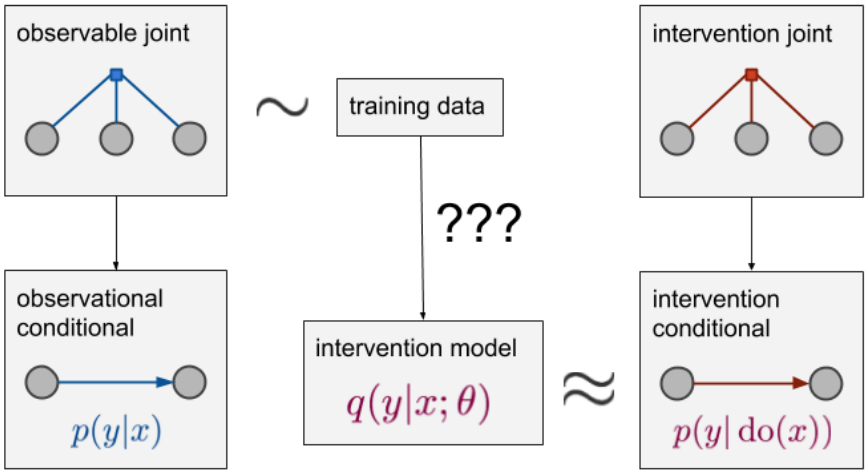
Bottom-line: a full causal model is a form of prior knowledge that you have to add to your analysis in order to get answers to causal questions without actually carrying out interventions. Reasoning with data alone won't be able to give you this.



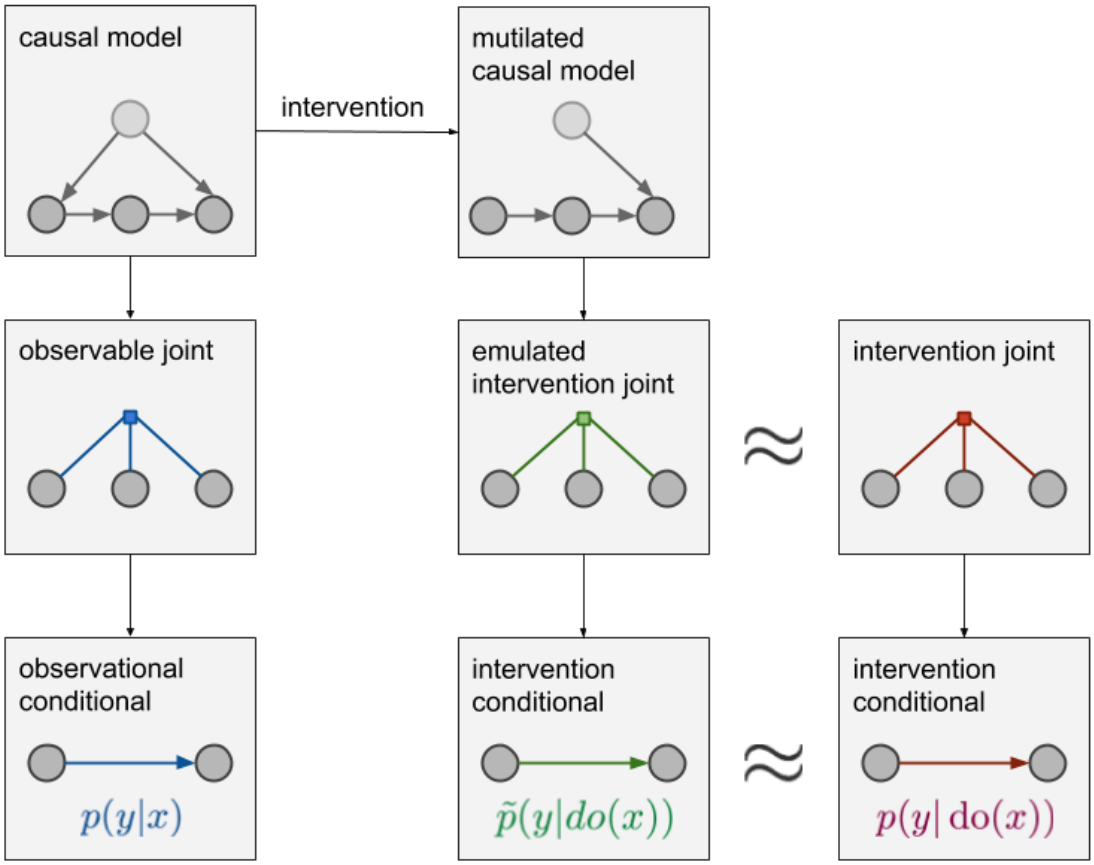
- The circles are RVs, the blue square is the joint probability distribution of the variables that it is connected to.
- Data $\{x, y, z\}_n$ is sampled from the blue square (joint distribution $P(X, Y, Z)$).
- The correlational relationship between Y and X (ex: predict Y from X) is given by $P(Y|X)$. I have measurements of Z but it is not used to predict Y.
- $P(Y|X) = \frac{P(X,Y)}{P(X)}$. Everything on the RHS can be obtained from $P(X, Y, Z)$.
- Using train data, I can estimate P(Y|X) using a parametric method (with parameters θ) such as linear regression.



Now, what if we're actually interested in $p(y|do(x))$ rather than $p(y|x)$? This is what it looks like:



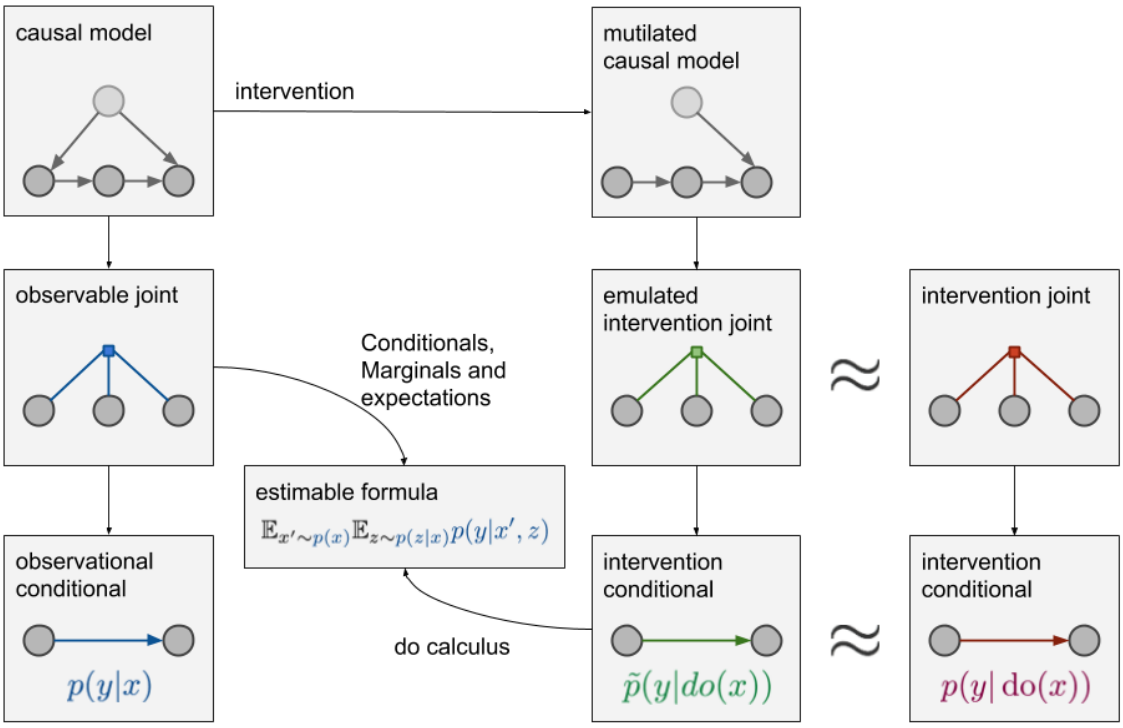
- $p(y|do(X))$ is the causal effect of Y from X. It is obtained from thhe intervention joint probability distribution $P_{do(X=x)}(x, y, z)$.
- If we cannot perform RCT, we are essentially trying to estimate $P(y|do(X))$ using $P(X, Y, Z)$



- From theory, I specify/assume the causal relationships and direction between the three RVs. This my causal model.
- Armed with these assumptions, I get the "emulated" intervention joint probability, from which I can approximate the intervention conditional probability: $\hat{P}(y|do(x)) \doteq P(y|do(x))$
- My estimated interventional conditional will equal to the actual interventional conditional if I get all my arrow direction and nodes correct in my causal model.

- To bridge between the blue and red stuff (through the green stuff), I need
  - Observational data
  - Assumptions (causal model based on how I think the world works)
- **Do-calculus** allows us to massage the green conditional distribution until we can express it in terms of various marginals, conditionals and expectations under the blue distribution.
- Do-calculus extends our toolkit of working with conditional probability distributions with four additional rules we can apply to conditional distributions with the **do** operators in them. These rules take into account properties of the causal diagram. Check out this paper: https://arxiv.org/abs/1305.5506

- Ideally, as a result of a do-calculus derivation you end up with an equivalent formula for $\hat{P}(Y|do(X))$ which no longer has any do operators in them, so you estimate it from observational data alone. If this is the case we say that the causal query $\hat{P}(Y|do(X))$ is *identifiable*. Conversely, if this is not possible, no matter how hard we try applying do-calculus, we call the causal query *non-identifiable*, which means that we won't be able to estimate it from the data we have.



- The "estimable formula" is the equivalent expression (rewriting) for the $\hat{P}(Y|do(X))$
- I need to observe X,Y, and Z in order to obtain $\hat{P}(Y|do(X))$ from do-calculus. Notice that Z must be observed even though we are only interested in causal relationship between X and Y.
- You can never fully verify the validity and completeness of your causal diagram based on observed data alone. However, there are certain aspects of the causal model which are empirically testable.
- In particular, the causal diagram implies certain conditional independence or dependence relationships between sets of variables. These dependencies or independencies can be empirically tested, and if they are not present in the data, that is an indication that your causal model is wrong.
- **Causal discovery** is an approach for coming up with the causal model empirically from data (instead of handcrafted from theory).